
DISTRIBUTED CONTROL WITH COLLECTIVE INTELLIGENCE

David H. Wolpert
NASA Ames Research Center
Moffett Field, CA 94035
dhw@ptolemy.arc.nasa.gov

Kevin R. Wheeler
NASA Ames Research Center
Moffett Field, CA 94035
kwheeler@mail.arc.nasa.gov

Kagan Tumer
NASA Ames Research Center
Moffett Field, CA 94035
kagan@ptolemy.arc.nasa.gov

Abstract

We consider systems of interacting reinforcement learning (RL) algorithms that do not “work at cross purposes”, in that their collective behavior maximizes a global utility function. We call such systems Collective INtelligences (COINs). We present the theory of designing COINs. Then we present experiments validating that theory in the context of two distributed control problems: We show that COINs perform near-optimally in a difficult variant of Arthur’s bar problem [Arthur] (and in particular avoid the tragedy of the commons for that problem), and we also illustrate optimal performance in the master-slave problem.

1 INTRODUCTION

Collective INtelligences (COINs) are (usually) very large, sparsely connected recurrent neural nets, whose “neurons” are themselves reinforcement learning (RL) algorithms. The distinguishing feature of a COIN is that its dynamics involves no centralized control, but only the collective effects of the individual neurons each modifying their behavior via their (local) RL algorithms. This restriction holds even though the goal of the COIN concerns the system’s global behavior.

One naturally-occurring COIN is a human economy, where the “neurons” consist of

individual humans trying to maximize their reward, and the "goal", for example, can be viewed as having the overall system achieve high gross domestic product. The robustness and power of naturally-occurring COINs suggests it would be useful to construct Artificial COINs (ACOINs), for example as controllers of distributed systems. This paper presents a preliminary investigation of this idea in the context of distributed control.

In designing an ACOIN, we start with a global utility function concerning global behavior. Our task is to initialize and then update the neurons' utility functions so that as the neurons improve their utilities, global utility also improves. (We may also wish to update the local topology of the net to the same end.) In particular, we need to ensure that the neurons do not "frustrate" each other as they attempt to increase their utilities. All of this must be done without centralized control.

We refer to the RL algorithms at each neuron as *microlearning*, and the updating of the ACOIN is known as *macrolearning*. For robustness and breadth of applicability, we assume essentially no knowledge concerning the dynamics of the full system, i.e., the macrolearning and/or microlearning must "learn" that dynamics, implicitly or otherwise. This rules out any approach that models the full system.

The problem of designing an ACOIN has never previously been addressed in full — hence the need for introducing new concepts like those described below. Nonetheless, this problem is related to previous work in many fields: distributed artificial intelligence, multi-agent systems, computational ecologies, adaptive control, computational economics, game theory (and in particular stochastic game theory, evolutionary game theory, and game-theoretic mechanism design), voting and auction theory, computational markets, iterated function systems, active walker models, self-organized criticality, spin glasses, reinforcement learning, recurrent neural nets, Markov decision process theory, network theory, and ant-based optimization.

WE SHOULD PROBABLY NOT CROSS REFERENCE SO THAT TWO PAPERS WILL NOT BE COMPARED FOR SIMILARITY!

In an associated paper [ngi], we design an ACOIN for distributed internet routing, achieving performance better than that of all previous RL-based approaches to internet routing [Littman, OTHERS]. In this paper we instead concentrate on improving our understanding of COINs. We do this by using computer experiments to investigate the theory of COINs in the context of two abstract distributed control problems: a more challenging variant of Arthur's bar problem in economics [Arthur] and the master-slave problem. In the context of these problems we validate some of the predictions of the theory of COINs. In particular, we confirm the necessity of "constraint alignment" for having the use of "wonderful life" neuronal utility functions result in optimal global performance: When there is constraint-alignment, the two COINs achieve 99% of the optimal for the bar problem and 100% of the optimal for the master-slave problem. In contrast, natural non-COIN-based utility functions encounter difficulties like tragedy of the commons for these two control problems, resulting in performance levels of at best 24.6% and 9% for the two problems investigated, respectively.

The next section presents an overview of those aspects of the mathematics behind ACOINs that are investigated in this paper. The following section presents computer experiments concerning that mathematics in the context of distributed

control. The final section presents conclusions and summarizes future research directions.

2 MATHEMATICS OF COINS

The mathematical framework for COINs in general is quite extensive [Us]. This paper concentrates on four of the concepts from that framework: subworlds, factored systems, constraint-alignment, and the wonderful-life utility function.

i) We consider the state of the system across a set of $T + 1$ consecutive timesteps, which we write without loss of generality as $t \in [0, \dots, T]$. Let $\underline{\zeta}_{\eta,t}$ be the (Euclidean-vector-valued) state of neuron η at time t . World utility is a function of the state of all neurons across all t , which we write as $G(\underline{\zeta})$.

ii) To simplify the design of the ACOIN, we group neurons into *subworlds*, all sharing the same *subworld utility*. In the ACOINs considered here, subworlds have no overlap, and a neuron's microlearner receives the utility of that neuron's subworld as its reward signal. We indicate the state of neuron σ in subworld ω at time t by $\underline{\zeta}_{\omega,\sigma,t}$, and by $g_{\omega}(\underline{\zeta})$ we indicate the subworld utility of subworld ω .

iii) One of the primary desiderata in designing an ACOIN is to have it be *subworld-factored*. This means that a change at time 0 to the state of the neurons in subworld ω results in an increased value for $g_{\omega}(\underline{\zeta})$ iff it results in an increased value for $G(\underline{\zeta})$.

For a subworld-factored system, ω 's increasing its own utility does not have "side-effects" on the rest of the system that decrease world utility. For these systems, the separate neurons pursuing their separate goals do not frustrate each other, as far as world utility is concerned. In particular, consider a Nash equilibrium of the system, i.e., a state in which no neuron can increase its utility by changing its action, given the actions of the other neurons [Nash, GT]. The following is proven in [us]:

Theorem 1: If a system is subworld-factored, then any state at time 0 of the system that constitutes a Nash equilibrium for the subworld utilities is a critical point of world utility. Conversely, a maximum of world utility is such a Nash equilibrium.

In other words, for a subworld-factored system, although it may be maximized at other points as well, world utility is assuredly maximized when each of the neurons' reinforcement learning algorithms are performing as well as they can, given each others' behavior. So we have the correct behavior if and when the learners reach an equilibrium. There can be no tragedy of the commons [totc], etc.

iv) A (perfectly) *constraint-aligned* system is one in which any change to the state of the neurons in subworld ω at time 0 will have no effects on the states of neurons outside of ω at times later than 0.

v) Let $Zero_{\omega}(\underline{\zeta})$ be defined as the vector $\underline{\zeta}$ modified by setting the values of all neurons in subworld ω across all time to 0. The *wonderful life subworld utility* (WLU) is $g_{\omega}(\underline{\zeta}) \equiv G(\underline{\zeta}) - G(Zero_{\omega}(\underline{\zeta}))$.

Intuitively, the WLU is analogous to the change in world utility that would have arisen if subworld ω had never existed. (Hence its name - cf. the Frank Capra movie.) Note though that no assumption is made that $Zero_{\omega}(\underline{\zeta})$ is consistent with the dynamics of the system. In particular, to evaluate the WLU we do *not* have

to infer how the system would have evolved if all neurons in ω were set to 0 at time 0. So long as we know ζ (by observation, in the process of macrolearning that occurs at time T), and so long as we know G (set in the problem specification of the ACOIN), we can evaluate the WLU. This is true even if we know nothing of the dynamics of the system. This dynamics-independence is a major strength of the WLU.

In addition to assuring the correct equilibrium behavior, there exist many other theoretical advantages to having a system be subworld-factored [us]. In this paper we use computer experiments to establish the applicability of these advantages to PROBLEMS. In particular, the following theorem, proved in [us], serves as the basis for the investigation of this paper:

Theorem 2: A constraint-aligned system with wonderful life subworld utilities is subworld-factored.

3 EXPERIMENTS

Two sets of experiments were conducted to investigate the theory behind ACOINs. The first was a variant of the bar-attendance model [Arthur, 94], the second was a more complicated model involving master/slave relationships.

Arthur [Arthur, 94] proposed a 'bar-attendance' model of economic market behavior which contained heterogeneous agents. This model has proven to be relatively easy to optimize [REFERENCE!!!]. To make the problem more challenging, the following variant was investigated.

The problem consists of N patrons that each pick one of 7 days to attend the bar on the following week. The bar has a fixed capacity. Each patron via reinforcement learning must predict which night to attend on the following week based upon past rewards. The global behavior was to have $G(x) = \sum_{k=1}^7 x_k \exp -x_k/c$ optimized given the capacity constraint c , where x is the attendance on one day. Three types of rewards were explored: 1. Tragedy-Of-the-Commons (TOC), 2. Global, and 3. Wonderful Life Utility.

The Tragedy-of-the-commons [Crowe, 69] used a local utility of $g(x) = G(x)/x$. The Global used $g(x) = G(x)$. This function resulted in everything being subworld factored. The wonderful life utility used $g_\omega(x) = G(x) - G(\text{Zero}_\omega(x))$. This function was close to constraint alignment and close to being subworld factored. The days in the week were not always treated equally. In one case only the middle day of the week was used in the utility calculations. In the other case all days were treated equally. The convergence rate of the three utility functions is compared. The resulting bar attendances are also compared.

The intent of the Master-slave experiments was to demonstrate how a poor choice in subworld memberships can adversely affect the Wonderful Life utility by the non-alignment of constraints (see Theorem 2). The bar problem was modified to resemble society where the actions of one member have an effect on other members of society. In the Master-slave experiments the capacity constraint on the bar was removed. Patrons were assigned to belong to clubs, where each club had at least one master, and each master had two slaves. Patrons were rewarded according to how many other patrons from the same club were also present at the bar on the same night. Slave patrons were required to take the same action as their master

patron.

The same set of utility functions were investigated for these Master-slave experiments. For the wonderful life utility, the closeness to constraint alignment depended upon how the subworld memberships were chosen. The subworld membership that resulted in the most constraint alignment was when members of the same club belonged to the same subworld. The worst case was when members of the same club were all in different subworlds. The third type of subworld membership explored was to assign the membership randomly.

4 RESULTS

Figures 1 and 2 show the convergence properties of the global performance in the bar problem for the three utility functions. In both cases the Wonderful Life utility converged faster to a better solution. The global utility would eventually achieve the same optimum but is very slow to converge. Figure 3 depicts the average deviation from the optimal daily attendance solution for the three utilities.

Table 1 shows the mean performance for the Master-slave problem for different subworld memberships using the Wonderful Life utility, and the performance of the TOC and Global utilities. The Global utility always had constraint alignment and was therefore able to achieve optimum. The Wonderful Life utility had perfect alignment when the subworld memberships correspond to the club membership.

Percentage of cases when random subworld membership assignment was able to surpass TOC in performance varied with the number of clubs as shown in Table 2.

References

- W. B. Arthur, *Amer. Econ. Rev.*, vol. 84, no. 406, 1994.
- B. L. Crowe, "The Tragedy of the Commons Revisited," *Science*, v.166, pp. 1103 - 1107, 1969.
- G. Galdarelli, M. Marsili and Y. C. Zhang, *Europhysy. Lett.*, vol. 40, pp. 479, 1997.
- J. F. Nash, "Equilibrium Points in N -Person Games," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 36, no. 48-49, 1950.
- M. Osborne and A. Rubenstein, *A Course in Game Theory*, MIT Press, Cambridge, MA, 1994.
- Us (1998)., (maybe ngi, Littman, others)

Table 1: Mean performance of Master-Slave Problem

Utility Type	Subworld Assignment	Mean Performance	No. Clubs
Wonderful	Club	0	2
Wonderful	Worse	-12	2
Wonderful	Random	-8.68	2
TOC		-1.16	2
Global		0	2
Wonderful	Club	0	3
Wonderful	Worse	-18	3
Wonderful	Random	-15.58	3
TOC		-1.75	3
Global		0	3
Wonderful	Club	0	4
Wonderful	Worse	-24	4
Wonderful	Random	-21.76	4
TOC		-2.34	4
Global		0	4

Table 2: Percentage of random subworld membership better than Tragedy of the Commons

No. of Clubs	Percent Better
2	21%
3	2%
4	0.08%

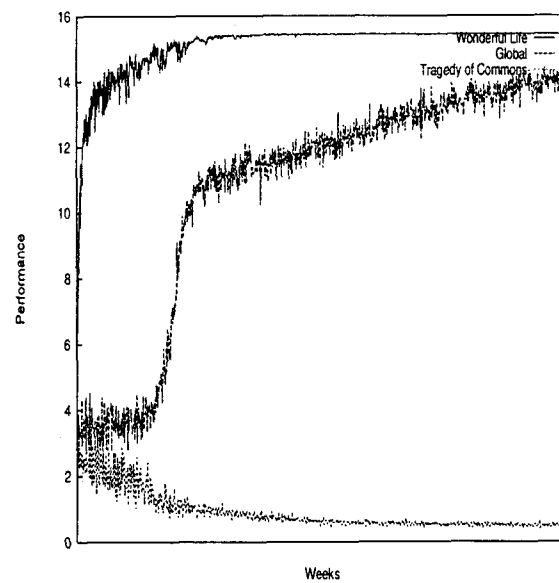


Figure 1. Average convergence when only middle day matters.

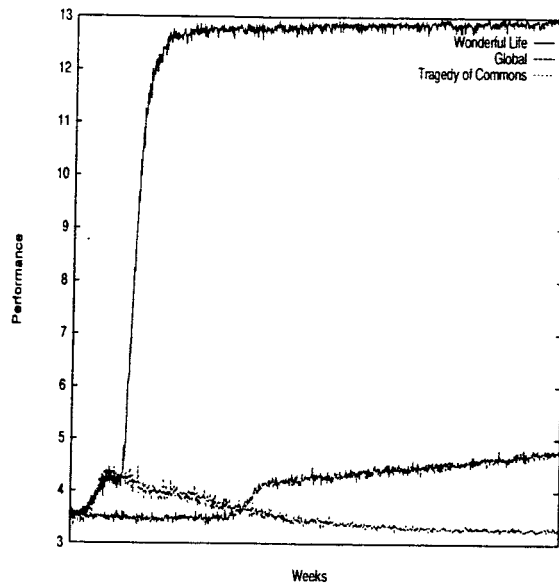


Figure 2. Average convergence with all days equal.

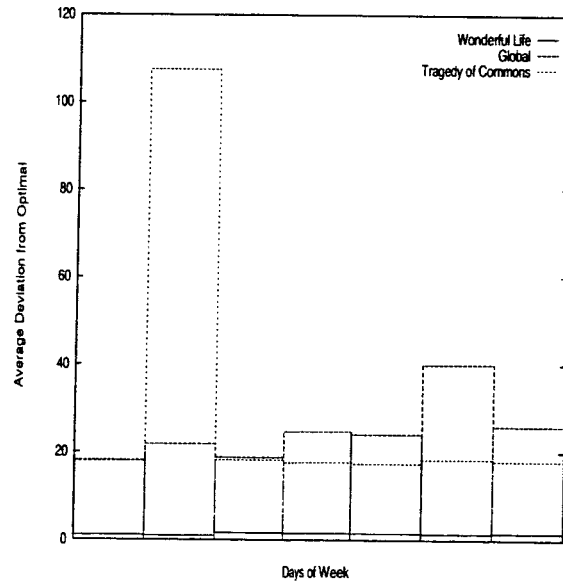


Figure 3. Average attendance deviation with all days equal.